DOCUMENT RESUME

ED 100 706 SE 018 646

AUTHOR Knisley, David B.; And Others
TITLE A Reliable Instrument for Participant Assessment of NSF Science Education Programs.
PUB DATE 74
NOTE 22p.

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS Curriculum Evaluation; Educational Research; Evaluation; *Inservice Teacher Education; Institutes (Training Programs); Participant Satisfaction; *Program Evaluation; Science Education; *Science Institutes; *Summer Institutes; *Teacher Education
IDENTIFIERS National Science Foundation; NSF; Research Reports

ABSTRACT

This document presents a study related to teacher education programs in general, but with specific attention directed to, National Science Foundation (NSF) institute programs for in-service teachers. The study was designed to answer questions pertinent to an assessment instrument designed and used with the Ball State University NSF institute programs. The reliability of the instrument, the extent of participant-perceived change in classroom emphasis given to 57 instructional topics included in the instrument, and whether a significant difference existed between the shift of emphasis of the 1972 institute participants and that of the 1973 participants were determined. All of the 1973 members took part in a pre-institute assessment and in two post-institute follow-ups. One follow-up test was administered on the last day of the institute and one the following spring. Changes in instructional emphasis were evaluated by the sign test. Comparison of mean growth increments was accomplished using a t-test. Analyses showed that participant-perceived increases in the level of emphasis were significant. No significant differences were found between the mean growth increments of the two groups. Instrument reliability was established, using analysis of variance in a modified intraclass correlation formula, as .93 and .97 for the 1972 and 1973 groups, respectively. (Author/EB)

ED 100706

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

BEST COPY AVAILABLE

A Reliable Instrument for Participant Assessment of NSF Science Education Programs

by

David B. Knisley
Biology Instructor
Northside High School
Muncie, Indiana 47304

Thomas R. Mertens
Professor of Biology
Ball State University
Muncie, Indiana 47306

Jon R. Hendrix
Coordinator School Science Visitation
Ball State University
Muncie, Indiana 47306

SE 018 646

BEST COPY AVAILABLE

For nearly two decades, the National Science Foundation funded institute programs that brought secondary school teachers of science and mathematics back to the college or university campus for further study. These institute programs began at the University of Washington in 1954, and by 1965 numbered 449 nation-wide. [11] Many of these institutes were designed to update the basic subject matter competence of teachers and to familiarize them with, and prepare them to teach, the courses produced by the various science curricular studies. The major intent of most institutes was to update secondary school science and mathematics teachers in both science content and instructional methodology.

As NSF policy changed in the early 1970s and as financial support for institute programs began to decline, institute directors and other concerned educators began to search for mechanisms to assess the effectiveness of institute programs. Numerous articles summarize studies which were designed to assess the effectiveness of NSF institutes. [1, 3, 7, 8, 9, 10, 12, 13] Evidence from these studies suggests that NSF-funded institutes have been influential in improving teacher competence and in fulfilling the objectives set forth by NSF. However, most of this evidence is of a subjective nature, usually obtained through the use of a questionnaire in which the participants simply responded to questions that pertained to how, or how much, the institute helped them.

The biology institute faculty at Ball State University has been of the opinion for some time that an institute evaluation instrument was needed that would provide data of a nature that could be analyzed statistically. A first

BEST COPY AVAILABLE

step pertaining to statistical analysis of data concerning teacher perceptions of institute effectiveness was taken by Hendren, Mertens and Nisbet. [8] This study was initiated in the spring of 1972 when a pre-institute assessment of the level of emphasis given to each of 55 instructional topics was made by each teacher selected to participate in the 1972 summer institute at Ball State University. This was followed by a post-institute assessment of the same 55 instructional topics and a determination of the amount of participant growth that had occurred with respect to these topics. The data collected through the pre-institute assessment provided a baseline with which comparisons were made concerning teaching emphasis given these topics after the institute. The establishment of these baseline data made a follow-up study amenable to statistical analysis.

By way of contrast, most institutes have been evaluated by post-institute questionnaires only. Such studies produced data which were not easily subjected to statistical analysis because no baseline data were available for comparison. The study by Hendren et al. [8] produced evidence of the change in emphasis given by participants to 55 instructional topics and did much to document institute effectiveness as perceived by the participants themselves. At least one serious doubt remained, however: "Was the assessment instrument reliable? Did the assessment instrument measure accurately what it was intended to measure?"

Unanswered Questions. As preparations for the 1973 biology institute at Ball State University were being finalized, the decision was made to pre/post assess the participants in a manner similar to that used with the 1972 institute participants. The assessment instrument was modified slightly

BEST COPY AVAILABLE

(e.g., two additional items were added, making a total of 57), and this new version was administered to the participants in the 1973 institute. The study summarized here was designed to answer three questions that are pertinent to the assessment instrument and to the Ball State University NSF institute programs.

The first concern of those administering the assessment instrument was the answer to the question, "Is the assessment instrument accurately measuring what it was intended to measure? Does the instrument effectively communicate to the participants so that valid interpretations of assessment results can be made?" Determination of the reliability of the assessment instrument became the primary task of this study.

The second concern of this study was to determine the extent of participant-perceived change in classroom emphasis given to the 57 instructional topics included in the assessment instrument. Specifically, the answer to the following question was sought: "Has statistically significant change in emphasis taken place following participation in the institute program?"

The third goal of the study was to determine whether or not a statistically significant difference existed between the shift of emphasis of the 1972 institute participants and that of the 1973 participants. The third major question to be answered was, "Did one group of participants change its emphasis more or less than the other group as measured by the assessment instrument?"

Methods. All of the 1973 institute participants took part in a pre-institute assessment and a post-institute follow-up using a modified form of the assessment instrument employed in 1972. In April 1973, before the institute

BEST COPY AVAILABLE

began, each of the 40 participants was asked to assess the emphasis he/she had placed on each of 57 instructional topics during academic year 1972-73. These topics included relevant teaching methodology as well as current developments in biological science. Each participant assessed his current level of emphasis, his desired level of emphasis, and the significance of each of the 57 topics for his students on a scale of 1-7. (See Table 1 for interpretation of the assessment scale.) It was anticipated that "desired emphasis" and "significance for students" would show a strong positive correlation and perhaps serve as a cross check within the assessment instrument. Surely, a teacher who thought that a particular topic was significant to students would desire to emphasize that topic in the classroom. A comparison of columns B and C in Table 2 reveals that the data generally confirm this prediction. The data collected through the pre-institute assessment (Table 2, column A) provided a baseline with which comparisons may be made concerning teaching emphasis given these topics after the institute.

The same assessment form was again administered to the participants on the last day of the summer institute program and the participant was asked to indicate the level of emphasis he/she desired to place on each of the instructional topics during 1973-74. Finally, the assessment form was mailed to each of the 40 participants in the spring of 1974, one year after the pre-institute assessment and seven months after the close of the summer institute. The same 57 topics were again assessed with respect to the level of emphasis actually given to the topic in the 1973-74 school year. Table 2 summarizes the findings of this study. The data obtained from the 1973-74

administration of the assessment form and from the use of the similar form in 1972 [8] provide the basis for the remainder of this article.

Data and Discussion

Reliability of the Assessment Instrument. Obviously the data obtained in a study such as has been described, are meaningful only if the assessment instrument can be demonstrated to have a reasonable measure of reliability. "Reliability" in this sense is simply "how accurately [the device] measures whatever it does measure." [14, p. 177] One approach to determining reliability is to administer the assessment instrument on several occasions to determine whether it performs similarly on repeated trials. Establishment of reliability by repeated performance is statistically valid when the resulting data constitute a ranking of alternatives (e.g., right or wrong responses). Clearly, responses on a sliding scale assessment instrument such as was used in this study, do not provide data of the requisite ranking type.

In the present study, reliability must be established for ratings rather than for rankings. For rating data it is possible to determine internal reliability; i.e., reliability may be established by using an analysis of variance in a modified intraclass correlation formula. [4, 6] In its simplest form this involves an analysis of variance and the calculation of the reliability coefficient, r, for all raters (participants completing the assessment), where

$$r = \frac{\text{mean square (all items)} - \text{mean square (error)}}{\text{mean square (all items)}} .$$

The value of r may range from -1.0 to 1.0 with the values between 0 and 1.0

indicating positive correlations between the raters' mean ratings.

Table 3 summarizes the analysis of variance statistics and reliability coefficents calculated for the post-institute assessment results for both the 1972 and 1973 participants. For the 1972 participants, for example, 30 raters (participants) provided complete data for the 55 items assessed by the instrument. In the case of the 1973 participants 33 raters each assessed the same 55 items (the two new items - numbers 18 and 27 in Table 2 were not included in the analysis). Using the data for the 1973 participants as an illustration, it may be inferred that if the 33 ratings for each of the 55 items were averaged and if we could correlate these averages with a similar set of averages from a comparable group of participants, the result would be about 0.97. Thus, the extremely high reliability coefficients, .93 and .97, provide positive evidence that the assessment instrument is truly "measuring accurately what it is intended to measure." The data obtained as a result of administering this instrument may be interpreted with a great deal of confidence, since a reliability of this magnitude exists.

Participant growth in 1973-74. The second goal of this study was the evaluation and interpretation of data obtained by the use of the assessment instrument with the 1973 institute participants. With the reliability of the instrument clearly established, one can have confidence in the data obtained by the administration of the instrument.

The mean assessment values reported in Table 2 suggest that, as perceived by the participants, the institute has been effective in stimulating an increase in the level of emphasis given to the 57 instructional topics. The data for

BEST COPY AVAILABLE

each of the topics were analyzed using a test designed to determine whether the increase in the level of emphasis was, in fact, statistically significant. The test used for this purpose, called the sign test, "is based on the signs of the differences (whether they are positive or negative) ignoring their magnitudes." [5, p. 295]

An example of how the data were subjected to the sign test follows: For item 27, "societal problems resulting from over population and mis-use of technology," 24 of the participants increased the level of emphasis in their teaching during academic year 1973-74, 3 decreased their emphasis, 7 did not change, and the data were incomplete for 2 of the participants. Occasional omissions on the completed assessment forms account for instances of incomplete data. In addition, four participants did not complete the post-institute assessment (spring 1974), thus reducing the total population studied to 36. Using the data for item 27, the null hypothesis, that the institute had no effect on determining the level of instructional emphasis, was tested. This is equivalent to testing the hypothesis that a positive change and a negative change were equally probable; that is, the chance of getting an increased level of emphasis (positive change) "is $p = 0.50$ against the one-sided alternative that $p > 0.50$." [5, p. 296] For this purpose the statistic z was calculated as follows:

$$z = \frac{x - \mu}{\sigma}$$

where x = the number of positive changes = 24; $\mu = np$, where n = total number of changes (both positive and negative) = 24 + 3 = 27, and $p = 0.50$; and $\sigma = \sqrt{np(1-p)} = \sqrt{27(0.50)(0.50)}$, because $p = 1-p = 0.50$.

Therefore, $z = \frac{24-13.5}{\sqrt{27(0.50)(0.50)}} = \frac{10.5}{\sqrt{6.75}} = 4.04$.

Using a table of z values [5], significance at the 1% level may be determined as $z = 2.33$. Hence, in the case of this topic, "societal problems resulting from over population and mis-use of technology," $z = 4.04$ is statistically significant at the 1% level and the null hypothesis is rejected. Therefore, for this topic, it may be concluded that the participant-perceived change in the level of emphasis following the institute is statistically significant. The reader will note that statistically significant increases (at the 1% level) in the degree of emphasis were given to 49 of the 57 instructional topics. Increases in emphasis were statistically significant at the 5% level for an additional six topics (items 7, 16, 19, 44 51, and 54). For only two items (21 and 38) among the 57 topics was there no evidence of statistically significant growth.

Since the type of institute assessment described herein was also employed with the participants in the 1972 summer institute, this study affords the opportunity to compare the two groups of participants and to obtain some evidence of the reliability of the assessment instrument in yet another way. For example, a comparison of Table 2 in this study with the comparable table in the study by Hendren et al. [8] reveals statistically significant increases (at the 1% level) with respect to 40 topics for <u>both</u> groups of participants. The increase in emphasis with respect to item 7 for both groups was statistically significant at the 5% level. Since the 1973 assessment form included two items (18 and 27) <u>not</u> included in the 1972 assessment, the two groups of participants gave similar responses to 41 of 55 topics included in both assessment instruments. These comparisons further increase confidence in the reliability of this technique of participant self-assessment.

BEST COPY AVAILABLE

When Hendren et al. [8] compared the "desired level of emphasis" before the institute (spring 1972) with the "desired level of emphasis" on the last day of the institute, they found that for 40 of the 55 topics the desired level of emphasis decreased following participation in the institute. This observation was attributed to a more realistic post-institute view on the part of the participating teachers as to what they would be able to accomplish upon returning to their respective classrooms. Such was not the case, however, for the 1973 participants (compare columns B and D in Table 2). For 37 of the 57 instructional topics in the 1973 assessment instrument, the participants increased the desired level of emphasis after participating in the institute. This seems to suggest that the institute was a motivating force in increasing participant interest in these institute topics.

Comparison of Growth Increments for Two Groups of Participants. The third goal of this study was to compare the participant-perceived changes in instructional emphasis between the 1972 and the 1973 groups of institute participants. Several observations suggest that the differences, if any, will be slight. These observations include: (1) the goals and objectives of the two institutes were similar; (2) the participants in the two institutes were all biology and/or life science teachers at the secondary school (7th-12th grades) level; (3) the assessment instruments used for the two groups of participants were identical with respect to the 55 items included in the analysis; (4) results of the sign tests, as stated above, indicate that growth was similar for many of the 55 items assessed; and (5) the reliability of the instrument has been clearly established.

In order to compare the two groups statistically, a standard two-tailed t test was applied to the null hypothesis that "no significant difference exists between the growth (change in instructional emphasis) of the two groups." In order to prepare the data for this test, the difference in the mean growth for each item from the time of pre-institute assessment until the time of post-institute assessment (actually seven months after the institute, as explained above) was calculated for each group of participants. The mean of these 55 differences (the two new items in the 1973 assessment form were discounted) was then calculated for each group. All statistics relevant to the t test are summarized in Table 4. The calculation of a value for t follows. Note that in both the explanation below and in Table 4, "group 1" refers to the 1972 participants and "group 2" refers to the 1973 participants.

If $\bar{X}_1$ = the mean for group 1 = 1.0836, $\bar{X}_2$ = the mean for group 2 = 1.2545, $\sum x_1^2$ = the sum of squared deviations for group 1 = 13.3952, $\sum x_2^2$ = the sum of squared deviations for group 2 = 14.5563, and n = 55 = the number of items assessed by each group, then according to Blommers and Lindquist [2],

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{\sum x_1^2 + \sum x_2^2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Substituting the numerical values in this formula gives:

$$\underline{t} = \frac{1.0836 - 1.2545}{\sqrt{\left(\frac{13.3952 + 14.5563}{55 + 55 - 2}\right)\left(\frac{1}{55} + \frac{1}{55}\right)}};$$

$$\text{then, } \underline{t} = \frac{-0.1709}{\sqrt{\left(\frac{27.9515}{108}\right)\left(\frac{2}{55}\right)}} = \frac{-0.1709}{\sqrt{.0094112}};$$

$$\text{finally, } \underline{t} = -1.7617.$$

Interpretation of this value of $\underline{t}$ using the appropriate number of degrees of freedom (108), reveals that the null hypothesis, that there is no significant difference between the mean growth increments of the two groups of participants, may be accepted since the probability associated with $\underline{t} = -1.7617$ is 8.09%. Thus, the difference ($\bar{X}_1 - \bar{X}_2 = -0.1709$) between the mean growth increment ($\bar{X}_1 = 1.0836$) of group 1 and the mean growth increment ($\bar{X}_2 = 1.2545$) of group 2 is <u>not</u> statistically significant. This finding provides a type of check on, and reinforcement of, the other statistical tests reported earlier in this article.

Summary

The implications of this study for the assessment of teacher education programs in general, and NSF science education programs in particular, would appear to be quite significant. The instrument used in pre/post assessment of the instructional emphasis given by in-service biology teachers to 55 topics yields data that are amenable to a number of kinds of statistical analysis. Instrument reliability was established using an analysis of variance in a modified intraclass correlation formula. Changes in instructional

emphasis with respect to each assessment item were evaluated by the sign test. Finally, comparison of mean growth increments for different groups of teachers was accomplished by using a t test.

The fact that the assessment instrument was shown to be highly reliable (r = 0.93 for 1972 data and r = 0.97 for 1973 data) lends credence to the sign test calculations which revealed that in the case of the 1973 participants, for example, the participant-perceived increases in the level of emphasis given to 49 of 57 instructional topics were statistically significant at the 1% level. The reliability of the instrument is further reinforced, since comparable results were obtained with two separate groups of participants. A statistically significant difference between the mean growth increments of the two groups of participants, as determined by the t test, was not found to exist.

Acknowledgment

The authors wish to acknowledge the invaluable assistance of Dr. Robert E. Hill, Director of the Examination Service at Ball State University, who identified the appropriate statistical tests and analyzed the data obtained in this study.

References

1. Anderson, Norman D. "Summer Institute in Earth Science at North Carolina State University." Journal of Geological Education 17 (5): 191-193; 1969.

2. Blommers, Paul, and E.F. Lindquist. Elementary Statistical Methods. Houghton-Mifflin Co. Boston, Massachusetts. 1960. pp. 346-349.

3. Dyche, Steven E. "NSF-trained Teachers Perform Better, 10-year Summer Institute Follow-up Shows." American Biology Teacher 36 (4): 236-238 + 244; 1974.

4. Ebel, Robert L. "Estimation of the Reliability of Ratings." Psychometrika 16: 407-424; 1951.

5. Freund, John E. Modern Elementary Statistics. Prentice-Hall, Inc. Englewood Cliffs, N.J. 1960. pp. 294-296.

6. Guilford, J.P., and Benjamin Fruchter. Fundamental Statistics in Psychology and Education, Fifth Ed. McGraw-Hill Book Co., New York. 1973. pp. 261-264.

7. Harrison, Anna J. "Science Education and the National Science Foundation." Journal of Chemical Education 48 (8): 492-493 + 514; 1971.

8. Hendren, Julianne, Thomas R. Mertens, and Jerry C. Nisbet. "A Study of an NSF Institute." American Biology Teacher 35 (9): 510-514; 1973.

9. Highwood, Joyce E., and Thomas R. Mertens. "Evaluations of NSF Summer Institutes." American Biology Teacher 34 (4): 215-221; 1972.

10. Kastrinos, William. "Summer Institute - A Follow-up." American Biology Teacher 29 (8): 620-621; 1967.

BEST COPY AVAILABLE

11. Kreighbaum, Hillier, and Hugh Rawson. An Investment in Knowledge. New York University Press, New York. 1969.

12. Ost, David H. "An Evaluation of an Institute for Teachers of Secondary School Biology." American Biology Teacher 33 (9): 546-548; 1971.

13. Thompson, John F., and William D. Romey et al. "An Evaluation of NSF-funded ESCP In-service Institutes." Journal of Geological Education 21 (5): 214-222; 1973.

14. Thorndike, Robert L., and Elizabeth Hagen. Measurement and Evaluation in Psychology and Education. John Wiley and Sons, Inc., New York. 1969. pp. 177-180.

Table 1. Rating scale used to assess each of 57 instructional topics.

1 = virtually no emphasis

2 = slight emphasis

3 = some emphasis, but below average

4 = average emphasis

5 = slightly above-average emphasis

6 = considerable emphasis

7 = high level of emphasis

THOMAS R. MERTENS

Table 2 . Mean assessment values for 1973 NSF institute participants. A, current emphasis, spring 1973 (before institute); B, desired emphasis, spring 1973 (before institute); C, perceived significance for students, spring 1973 (before institute); D, desired emphasis, August 1973 (end of institute); E, actual emphasis, spring 1974; F, sign test z value (** = significant at the 1% level; * = significant at the 5% level).

INSTITUTE TOPIC	A	B	C	D	E	F
Molecular biology						
1. Biologically significant molecules	3.86	4.89	4.49	4.80	4.71	2.75**
2. Functional groups	3.11	3.91	3.71	4.11	3.80	2.56**
3. Chemical constituents of cells	3.31	4.34	3.88	4.72	4.34	3.27**
4. Chromatography	1.94	3.44	3.18	4.25	3.12	3.54**
5. Electrophoresis	0.53	1.17	1.11	2.72	1.34	4.12**
6. Autoradiography	0.43	1.17	1.21	2.30	1.53	4.47**
Cell structure and function						
7. Cell organelles	4.77	5.36	5.00	5.26	5.20	2.13*
8. Cyclosis	2.54	3.22	2.97	4.37	4.41	3.80**
9. Photosynthesis	4.66	5.46	5.18	5.69	5.49	3.96**
10. Energy production	4.03	5.08	4.89	5.31	5.34	4.43**
11. Mitosis	4.46	5.33	4.89	5.09	5.34	3.13**
12. Meiosis	4.46	5.19	5.00	5.09	5.31	3.13**
Genetic biology						
13. Basic principles of Mendelian genetics	4.09	5.09	5.00	4.94	5.00	2.75**
14. Human sex chromatin	3.11	4.88	4.71	4.06	4.14	3.53**
15. Human chromosome aberrations	3.15	4.63	4.63	4.83	4.61	4.43**
16. *Drosophila* genetics	2.29	3.71	3.76	3.86	3.32	1.96*
17. *Sordaria* genetics	0.23	0.76	0.74	3.17	1.70	5.00**

THOMAS R. MERTENS

INSTITUTE TOPIC	A	B	C	D	E	F
Genetic biology						
18. Societal problems resulting from new genetics knowledge and technology	2.24	4.86	4.74	5.34	4.51	4.85**
Biologic diversity						
19. Evidences for evolution	3.54	4.64	4.37	4.81	4.51	2.20*
20. Mechanism of evolution	3.23	4.50	4.11	4.64	4.37	3.40**
21. Principles of biosystematics	3.63	4.31	4.09	4.22	3.83	0.63
Ecological principles and environmental problems						
22. Ecologic succession	3.69	4.81	4.60	4.86	4.29	2.65**
23. Competitive exclusion (Gause's principle)	1.58	2.50	2.43	4.39	3.12	3.65**
24. Problems of pollution	4.31	5.47	5.54	5.86	5.04	2.65**
25. Eutrophication and water quality	3.06	4.69	4.76	5.31	4.23	3.53**
26. Population growth curves	3.31	4.69	4.74	4.67	4.56	3.27**
27. Societal problems resulting from over population and mis-use of technology	3.26	5.14	5.29	5.31	4.89	4.04**
Philosophic basis for biology instruction						
28. Assessing the direction and significance of developments in biology training	3.35	4.94	4.82	4.63	4.91	3.53**
29. Evaluating teacher goals	3.56	5.03	4.70	5.57	5.75	4.95**
30. Writing performance objectives	3.56	4.97	4.91	6.00	5.58	4.62**
31. Considering the affective domain	2.94	4.27	4.16	5.57	5.47	4.38**
32. Considering the cognitive domain	3.39	4.18	4.00	5.54	5.41	3.78**
33. Evaluating the function of evaluation	2.79	4.09	3.88	5.03	5.31	4.54**
34. Constructing test items	4.06	5.31	5.15	5.94	5.81	3.89**

THOMAS R. MERTENS

BEST COPY AVAILABLE

INSTITUTE TOPIC	A	B	C	D	E	F
Interpersonal challenges for the biology teacher						
35. Considering characteristics of the effective teacher	3.66	5.33	5.21	5.78	5.43	2.89**
36. Assessing the significance of self-concept	3.47	4.74	4.70	5.19	4.88	3.27**
37. Enlisting administrative support	3.91	5.14	5.18	5.17	5.34	4.20**
38. Working with peers	4.97	6.00	5.89	6.08	5.63	1.63
Curricular materials for biology instruction						
39. Teaching BSCS standard courses	3.00	4.06	3.91	3.61	3.91	2.40**
40. Using BSCS lab blocks	1.18	2.62	2.55	3.89	2.79	3.53**
41. Working with second level BSCS materials	0.62	1.30	1.27	1.56	1.97	3.96**
42. Working with BSCS special materials	0.94	2.08	2.26	2.92	2.44	4.81**
43. Developing teaching units for local use	1.66	3.72	3.83	4.75	4.00	4.27**
Teaching strategies						
44. Developing audiotutorial materials	2.46	4.67	4.71	3.61	3.36	1.73*
45. Developing electronic-response materials	0.69	1.97	1.94	1.72	1.11	2.68**
46. Teaching through inquiry	3.97	5.69	5.57	5.89	5.40	3.14**
47. Assessing contract learning	1.71	3.71	3.58	3.97	2.94	2.71**
48. Implementing modular scheduling	1.45	3.03	3.06	3.03	2.39	3.41**
49. Teaching controversial topics	3.03	4.49	4.56	5.14	4.86	4.27**
50. Experiencing microteaching	1.29	2.29	1.91	3.47	2.35	3.02**
51. Using TV in biology instruction	1.62	3.26	3.33	2.78	2.38	2.24*

THOMAS R. MERTENS

INSTITUTE TOPIC	A	B	C	D	E	F
Facilities, materials and resources for biology instruction						
52. Guidelines for the biology library	2.24	4.60	4.53	3.86	3.74	3.66**
53. Selecting equipment and facilities	3.79	5.06	5.06	4.47	5.03	3.40**
54. Designing biology laboratories	2.31	3.66	3.88	3.17	3.49	1.96*
55. Identifying sources of supplies and living materials	3.54	5.22	5.00	4.83	5.09	2.65**
56. Using outdoor education areas	2.94	5.61	5.69	5.14	4.00	2.45**
57. Employing community resources	2.57	5.33	5.37	5.00	4.14	4.04**

THOMAS R. MERTENS

Table 3. Internal reliability of assessment instruments was calculated using data obtained from administering the assessment in the spring following each institute.

A. Assessment form completed in spring 1973 by 1972 participants (30 countable)

Sources of variability	Sums of squares	degrees of freedom	mean square
Items	1596.7991	54	29.5704
Raters	1289.7384	29	44.4737
Error	3195.8009	1566	2.0407
Total	6082.3384	1649	

$$\text{reliability (all raters)} = \frac{29.5704-2.0407}{29.5704} = .9310$$

B. Assessment form completed in spring 1974 by 1973 participants (33 countable)

Sources of variability	Sums of squares	degrees of freedom	mean square
Items	3532.0313	54	65.4080
Raters	1048.0400	32	32.7512
Error	3865.9527	1728	2.2372
Total	8446.0240	1814	

$$\text{reliability (all raters)} = \frac{65.4080-2.2372}{65.4080} = .9658$$

THOMAS R. MERRENS

BEST COPY AVAILABLE

Table 4. Statistics obtained in analysis of data for calculating a *t* test for the comparison of the means of differences of pre/post institute assessment data for two groups of participants. The calculated value of *t* = -1.7617. The two-tailed probability level with 108 degrees of freedom = 8.09%.

Component of calculation	Group 1 (1972 participants)	Group 2 (1973 participants)
Mean of differences of 55 pre/post assessed items	1.0836	1.2545
Variance	.2435	.2646
Standard deviation	.4935	.5144
n (number of assessment items)	55.0000	55.0000
Sum of scores	59.6000	69.0000
Sum of scores squared	77.9800	101.1200
Sum of squared deviations	13.3952	14.5563

THOMAS R. MERTENS